Learning in Collective Dilemmas

Francisco Pedro Durão
Instituto Superior Técnico
Universidade de Lisboa
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
Email: francisco.durao@tecnico.ulisboa.pt

***Abstract*—Explaining the emergence of altruistic cooperation is a very important matter. By knowing how cooperation emerges we can create better environments for it to develop and be maintained. Altruistic cooperation may seem counterproductive for the individual but it is widespread in the animal world. While cooperation is traditionally studied in a two-person interaction - framed in the well-known Prisoners Dilemma - there are several examples of cooperative behavior in the form of collective dilemmas: from group hunting, communal activities in human settlements to agreements between countries to promote ecological sustainability. Most of these scenarios can be formulated in terms of a public goods game (PGG), which is the main focus of this thesis. Evolutionary Game Theory (EGT) provides a way for us to study the complex interactions within a population, assuming that individuals adapt through social learning. There has been a lot of research effort under EGT trying to explain how cooperation can beat selfish behavior in the real world. Notwithstanding, the effects of individual-based learning on cooperation under a PGG have not received much attention. Will cooperation emerge in a population where agents are trying to learn which behavior leads to the best possible outcome, based on their own experience? This is the question that propelled this work. We shed some light on how selfish and rational agents behave when given the choice to incur a cost to themselves to produce a benefit for another. As conflicting as these ideas may seem, it was shown that selfish agents can still choose to cooperate, under certain circumstances.**

***Index Terms*—Cooperation, Public Goods Games, Evolutionary Game Theory, Reinforcement Learning, Q-learning, Social Learning, Social Dilemma**

I. Introduction

Explaining the emergence of altruistic cooperation is not an easy task. Altruistic cooperation happens when an individual is willing to pay a cost for another to receive a benefit [1]. This way, altruistic cooperation (henceforth solely designated as cooperation) may appear to be irrational. Notwithstanding, cooperative behavior can be observed throughout many animal species. Lions, chimpanzees and African wild dogs cooperate in group hunts [2], [3], [4], some birds risk their lives to alert others that predators are near [5]. Human populations provide the best examples of such behavior: from communal activities in small villages [6] to international relations [7] and macroeconomic behavior [8], cooperation exists and understanding it is a challenging endeavor. It also happens at different levels of organization: some authors argue that cooperation was even in the origin of multicellular organisms [9].

	Cooperate	Defect
Cooperate	$benefit - cost$	$-cost$
Defect	$benefit$	0

TABLE I
Payoffs of Prisoner's Dilemma for the row player.

Cooperation seems to go against Darwin's principle of selection of the fittest. Selection of the fittest might have led us to believe that evolution is nothing but competition. Why would an individual risk its survival to provide a benefit to another? The duality that these concepts (evolution and cooperation) appear to have, impelled a lot of research effort, trying to understand the emergence of cooperation and why it creates better individuals and better societies.

A. Game Theory

In a nutshell, game theory represents the mathematics of conflicts of interest. This framework was first proposed by von Neumann and Morgenstern [10]. Its purpose is to study the strategic decisions and outcomes of rational agents when interacting with each other. One of the most famous example of conflict in Game Theory is the Prisoner's Dilemma. In this game, an agent can decide to give a benefit to other, incurring in a cost to himself. The payoffs earned by one player are summarized in Table 1.

In this type of games the best possible outcome for one agent is for him to defect while the other agent cooperates. This way he gets the benefit without the loss of the cost. Also, an agent that defects will have no incentive to change his behavior. Changing his behavior will always result in a worse payoff, regardless of what the other agent plays. This means that a game played by rational and selfish agents will end with both defecting, leading to a situation where both agents get 0 payoff. Individually an agent might think that defection is the best choice. However, for both agents cooperation is the choice that maximizes their payoffs. Because of this contradiction we are in presence of a **social dilemma** [11], where cooperation seems irrational but leads to the best outcome for the population as a whole.

We can extend a game from pairs to groups. The Public Goods Game (PGG) is a game similar to the Prisoner's Dilemma but played by a group of agents. In this game, each individual can contribute a quantity to a pile. That pile is then multiplied by a factor and distributed evenly by all the players.

Cooperation means contributing to the pile and defection is just to receive the division of the pile, without contributing. Again, rational players facing this dilemma will try to defect, because defectors always get a higher payoff in mixed groups. However, this reasoning ignores the collective (population-wide) dynamics, where a continuous process of behavioral revision takes place [12]. We can model this process using Evolutionary Game Theory where the agents payoffs are used to change their behaviors.

B. Evolutionary Game Theory

Darwinism, in short, explains evolution as a process where the best individuals are selected for reproduction more often than others. If we associate payoffs with fitness in the models described previously, the idea of Game Theory can be used to understand evolution and animal behavior. Evolutionary Game Theory (EGT) is able to represent frequency-dependent evolutionary processes, typical of natural selection, but also of cultural evolution, which occurs at faster time-scales. Cultural evolution is often based on social learning (see section II-B), which mathematically is equivalent to a common evolutionary process: instead of reproduction of the fittest, traits that offer higher payoffs are imitated more frequently. Unlike Game Theory, the fitness of an individual agent can not be measured in isolation; rather it has to be evaluated in the context of the full population in which it lives.

C. Problem

Given the abundance of competition and the way cooperation seems counterproductive for the individual, the main problem is to understand why cooperation is so widespread in the animal kingdom and in our own societies. **Will cooperation emerge in a population where agents are trying to learn which behavior leads to the best possible outcome, based on their own experience?** This is the main question we aim to answer. This work will focus on explaining how **individual learning** can influence the emergence of cooperation. Will the effect of individual learning be similar to that of social learning, traditionally studied under the framework of EGT? Specifically, we will focus on computational simulations of large populations, consisting of agents using **reinforcement learning** playing a Public Goods Game (PGG). We will focus on simulations because it is not clear how one can study the population dynamics analytically for Individual Learning. The branch of individual learning used will be reinforcement learning because the agents do not need to know the strategic nature of the game nor what other players are thinking. This makes them suitable and easy to implement for all types of games. Also, the agents' desires are easy to model: they simply want to obtain the highest possible payoff. This way we can observe if cooperation will emerge even when agents are behaving like selfish, rational individuals. PGGs provide a good model for some interactions present in the real world, as is explained in more detail on section III-A

This is a very important problem. The solutions may have an informative and orienting purpose. Informative as they may help explain why or how cooperation emerged in the course of animal evolution. Orienting because by knowing how cooperation emerges we can create better environments for it to develop and be maintained when needed. Global warming has been described as one of the greatest public goods dilemmas that humans face. Knowing which factors help the emergence of cooperation could help the cooperative agreements between countries to promote ecological sustainability [12], [13]. This knowledge can also help define policies that help the management of commons, surpassing the Tragedy of the Commons posed by Hardin [14].

II. Related Work

Computer simulations of artificial societies helped enlarge the concept of traditional Game Theory. A lot of work about the mechanisms by which cooperation is able to emerge in those artificial societies was already made. First we will present a set of mechanisms that promote cooperation in populations where individuals are playing Prisoner's Dilemma, the most common game to study cooperation. Here the game is played between two agents, although with Social Learning (see section II-B) this game is modeled using a whole group of agents interacting at the same time (a Public Goods Game (PGG)).

A. Mechanisms that favor cooperation

A lot of research effort has already been made, trying to find mechanisms that help cooperation beat defection in the struggle to be a stable strategy. Nowak summarizes five of these mechanisms that encourage individuals to cooperate [1]. They are kin selection, direct reciprocity, indirect reciprocity, structured populations and group selection. Each of this mechanisms is presented with more detail next:

Kin Selection: Kin Selection may be defined as the cooperation between relatives. Relatedness is defined as the probability of sharing a gene. It can be assumed that the purpose of one individual is the safeguard and propagation of his genes. In this case, by helping a related individual the payoff of the relate will contribute to the payoff of the individual, proportionally to their relatedness. Hamilton's rule [15] states that the coefficient of relatedness between two individuals must exceed the $cost/benefit$ ratio for cooperation to occur.

Direct Reciprocity: An agent might interact with another agent more than once. If this repeated interaction occurs, one might remember the behavior of the other and take that into account in future decisions. Trivers was pioneer in the study of reciprocity as the basis for cooperation [5]. In his research he observed examples of cooperation among animals that kin selection was not able to explain. This led Trivers to propose Direct Reciprocity as another mechanism to explain

cooperation.

Indirect Reciprocity: Instead of only helping others who can help us, indirect reciprocity is about helping those who can not help us back. Helping someone establishes a good reputation, which we hope will bring us future benefits. Nowak and Sigmund [16] proposed a model using EGT where all the agents in a population have an image score visible to all other agents. In that model two agents are randomly chosen, one being the potential donor of some altruistic act and the other being the recipient. If the donor chooses to cooperate he incurs a cost to himself while the recipient receives a benefit greater than that cost. Doing this increases the image score of the donor; conversely, if the donor chooses not to cooperate his image score is decreased. This way, agents that help others can be rewarded while those who do not are punished. It is shown that if the strategy used by one agent takes into account the other's image score, cooperation dominates the population.

Structured Populations: So far, all mechanisms presented use well-mixed populations where everybody interacts equally likely with everybody else. However, real populations are highly structured. Spatial structures or social networks make some individuals interact more often than others. To study if the emergence of cooperation is affected by the structure of the population one can use evolutionary graph theory [17]. A cooperator pays a cost for each neighbor to receive a benefit, whereas defectors pay no cost but their neighbors receive no benefits. This way, cooperators can prevail by forming network clusters, where they help each other. The resulting "network reciprocity" is a generalization of "spatial reciprocity", initially proposed by Nowak and May [18]. In this study the authors conclude that spatial structure seems to be crucial to the emergence of cooperation. However, saying that each person connects with an exact number of neighbors is an over-simplistic assumption. Recent advances in network theory [19], [20], show us that populations are organized in complex interaction structures, ranging from random-like graphs to scale-free networks. Later, Santos, Rodrigues and Pacheco [21] developed a model where Prisoner's Dilemma was played in a scale-free network, and it was concluded that cooperators can take advantage of this kind of structure, being cooperation the strategy that dominated the population.

Group Selection: What if we think not only about the individuals but also about the groups they take part in? A group of cooperators might be more successful than a group of defectors. A study done by Traulsen and Nowak [22] tackled this question. In this study, the population was divided into groups. Cooperators help others in their own group while defectors do not help. Individuals reproduce with a probability directly proportional to their payoff and their offspring are added to the same group. When the size of a group reaches a threshold, the group will split in two groups. In this case, another group becomes extinct to constrain the total population size. Although only individuals reproduce, selection emerges on two levels. On the lower level (within groups) selection favors defectors, while on the higher level (between groups) selection favors cooperation. The authors concluded that if the $benefit/cost$ ratio is sufficient, cooperation will spread.

B. Social Learning

Social learning happens when an individual imitates the behavior of another. To achieve this purpose,a pairwise comparison rule can be implemented. After playing a round of the Public Goods Game (PGG) agents receive their payoff. Then, a focal agent and one second agent are randomly selected from the population. The focal agent will imitate the behavior of the second agent with a probability proportional to the difference of fitness between those agents. Usually, the Fermi function is used as a pairwise comparison rule, as studied by Traulsen, Nowak and Pacheco [23]. The strategy of A will replace that of B with a probability given by the Fermi function:

$$pr = \frac{1}{1 + e^{-\beta(f_A - f_B)}} \tag{1}$$

The reverse will happen with probability $1 - pr$. When using this update rule, imitation will occur with probability proportional to the difference between the fitness of both individuals $(f_A - f_B)$. The parameter β controls the intensity of selection. If this value is 0, imitation will occur randomly and with probability equal to 0.5. If β is very large, imitation will strongly depend on the difference between payoffs.

This imitation process is equivalent to what we usually understand as evolution. We can view evolution not only as reproducing new agents, but also as reproducing ideas and making other agents willing to imitate one's behavior. This spreading of behavior is usually how evolution is modeled using EGT. As said on section I-A, in a PGG each individual can contribute a quantity to a pile. That pile is then multiplied by an enhancement factor and distributed evenly by all the players in the group.

Public Goods Game: Pacheco *et al.* [24] showed that if that enhancement factor is greater than the group size cooperation emerges. Conversely, whenever the enhancement factor is smaller than the group size cooperators stand no evolutionary chance. Then, they introduced a threshold for participants, below which no public good is produced. This means that if the number of individuals cooperating in a group is smaller than the threshold, that group produces no goods at all. This changes the game from the traditional PGG to a N-Person Stag Hunt game. In this game the dynamics of the population are more complex than what was seen with the PGG. The introduction of this threshold, in most situations, allowed for the emergence of cooperation even when the enhancement factor was smaller than the group size. They also showed the importance of the relation between group size and population size. Only for group sizes much

smaller than the population size can cooperation emerge. If the population is small or the group size spans nearly the entire population there can be observed a "spite" effect, first noted by Hamilton [25], which is detrimental for cooperation.

N-Person Snowdrift Game: On another study Souza, Pacheco and Santos [26] studied the population dynamics of another game: the N-Person Snowdrift. In this game there also exists a threshold below which no public good is produced. If this threshold is reached, a benefit is produced. Unlike the PGG, the benefit shared by individuals is always the same: it does not depend on the number of cooperators. Also unlike the PGG, the cost is divided by all contributors. If there are more agents willing to cooperate each of them has to pay less to produce the benefit. Similarly to the aforementioned study , Souza, Pacheco and Santos concluded that to provide the best conditions for the emergence of cooperation in the snowdrift game, the group size should be much smaller than the population size. The introduction of a threshold similar to the one employed by Pacheco et al. [24] also provided similar results. The population dynamics are more complex and under certain conditions this threshold provides an incentive to cooperate.

Structured populations: Similarly to what was previously said about Structured Populations (see section II-A), F. Santos, M. Santos and Pacheco [27] concluded that cooperation in Public Goods Games may emerge as an outcome of social diversity. In this study, the agents are organized in a scale-free network (see section II-A) and play only with their neighbors. The diversity in connectivity introduced with the scale-free network greatly helped the emergence of cooperation. This effect was already visible when each cooperator gave a fixed cost per each group he participated. But when each cooperator gave a fixed cost per individual (each cooperator contributes a cost equally divided between all groups he participates) this effect was amplified and the fraction of cooperators on the population was significantly higher. This means that social diversity is not the only important factor for the emergence of cooperation. The way each individual decides what amount to contribute also plays a big role.

Risk: The perception of risk can also play a role on the emergence of cooperation. Sometimes certain goals are shadowed by the uncertainty of its achievement. Milinski et al [13] illustrated this in actual experiments making use of a repeated game in which the perception of risk was shown to be a great factor when dealing, in that case, with the problem of climate change. Santos and Pacheco [12] modeled a similar problem using evolutionary game theory. The agents played a game similar to a PGG with threshold. The agents had an initial endowment. Cooperators contribute a fraction of their endowment, whereas Defectors do not. This time, however, if the group did not reach the threshold all agents would lose their remaining endowments according to a certain probability (the risk). Santos and Pacheco [12] arrived at a similar conclusion as Milinski et al [13]: decisions under high risk significantly raise the chances of coordinating actions to achieve a common goal.

C. Individual Learning

With Individual Learning (IL) agents can learn over time about the game or about the behaviors of others. In contrast with EGT, players use the history of the game to decide what action to take next. Some of the models that have been used for learning in game theory include reinforcement learning, myopic response, fictitious play, and rational learning [28]. These models are presented in ascending order of sophistication according to the amount of information agents use and their computational capabilities.

Reinforcement learning: When an agent repeatedly ends up and takes actions in the same situation, he can rely on his experience to choose or avoid certain actions based on their immediate consequences. This is the notion behind reinforcement learning [29]. Reinforcement learners only use the immediately received payoff to adjust the probability of conducting the same action accordingly. Actions that led to better outcomes in the past tend to be repeated in the future, whereas choices that led to unsatisfactory experiences are avoided. This way, reinforcement learners are unaware of the strategic nature of the game. Reinforcement learning can be implemented in different ways, for example, by using *Roth-Erev learning* [29] or *Q-learning* [30]. The latter being the focus of this work. A more detailed explanation of reinforcement learning is provided on section III-E.

Myopic Response: For this family of learning models, the agents need to have complete information about the game being played. This means that each player knows the payoff that he will receive in each possible outcome of the game. They also need to know the actions that every other player selected in the immediate past. Agents have a static and deterministic perception of the environment. This means that when an agent makes his next decision he assumes that every other agent will keep his current action unchanged; and that an agent can predict what future state he will be in by taking into account its current state and all actions taken by other agents. Working under such assumption, each agent can identify the set of strategies that would lead to an improvement of his current payoff. Because in this model agents assume their environment is static and deterministic, it is said that they respond in a myopic fashion: they ignore the implications of current choices on future choices and payoffs.

Fictitious Play: As with Myopic Response, players in Fictitious Play (FP) models are assumed to have a certain model of the situation and decide optimally on the basis of it. However, instead of assuming other agents will play the same action they did previously, a FP agent assumes that each of the other agents is playing a certain mixed strategy. The estimation of this mixed strategy is equal to the frequency

with which the counterpart has selected each of his available actions up until that moment. Thus, instead of considering the actions taken by every other player only in the immediate past, FP agents implicitly take into account the whole history of the game.

Rational learning: Kalai and Lehrer were pioneers in the study of rational learning [31]. This is the most sophisticated model of learning in IL. Agents in this model are assumed to be fully aware of the strategic context they are embedded in. They also have a set of subjective beliefs over the behavioral strategies of the other players. Agents must assign a strictly positive probability to any strategy profile that is coherent with the history of the game. This means that agents must be aware of all possible actions made by other agents. Finally, players are assumed to respond optimally to their beliefs with the objective of maximizing the flow of future payoffs.

III. Model

After introducing the basic concepts and reviewing related work, it was decided to propose a new model to study in what conditions cooperation can emerge. The proposal is a computer simulation of a Public Goods Game (PGG) played by a large population, where each individual uses reinforcement learning in an attempt to improve his payoffs. Individually, agents are selfish and rational. Considering this fact, it will be interesting to see under what conditions cooperation can beat defection in the struggle to be a stable strategy.

A. Public Goods Games with Thresholds

PGGs provide a good model for some interactions present in the real world. In a PGG, each individual may contribute a quantity to a pile. That pile is then multiplied by an enhancement factor and distributed evenly by all the players in the group. Cooperation means contributing to the pile and defection is just to receive the division of the pile, without contributing. The population is divided into groups. In each round each of those groups plays a PGG. This means that in each round an agent may play only once (if he is only in one group) or several times (if he takes part in multiple groups). Groups are formed randomly and the payoff of an individual agent is the average of all the payoffs he received in that round. This group sampling is explained in more detail in section III-B. In this type of model, it can make sense to introduce a threshold below which no public good is produced. Only when the number of cooperators surpasses this threshold does the group produce any benefit. Consider the way lionesses hunt in groups [2]. Two individuals are not enough for the cooperative hunt to be successful. At least three individuals are required for the group to catch prey and then the more individuals added, the more successful the group can be. This example of animal behavior could be modeled as a PGG with a threshold of three individuals.
Formally, this PGG is defined by some variables:

Payoff obtained	C	D
1≤k<M	$-c$	0
k≥M	$\frac{kFc}{N} - c$	$\frac{kFc}{N}$

TABLE II
Payoff values for the PGG for a group with k cooperators

- Each agent can be either a cooperator (C) or a defector (D) in each round. Players can change strategy between rounds.
- The variable F is the enhancement factor.
- c is the cost each contributor pays to the group.
- The variable N is the size of the groups playing.
- The variable Z is the population size.
- The variable M represents the threshold. For games without threshold this simply means $M = 1$.

Therefore $\frac{kFc}{N}$ represents the division of the pile. For a group with k cooperators the payoff matrix is presented in table II. It is worth noting that in any mixed group Cs are always worse off than Ds. This leads us to the conclusion that, under traditional Game Theory, everybody ends up defecting, thus foregoing the public good.

B. Group Sampling

The fitness of one agent can be measured by the average payoffs of all the games that agent played in that round, following what was done in Social Learning [24]. Groups are formed by selecting N agents randomly from the population of Z agents. This is called the *random matching model*. Agents are selected randomly and they are not able to identify each other. This way we ensure that they are not learning how to play against a specific individual. The population is well-mixed, which means that all players are equally likely to be selected to participate in a specific group. For Structured Populations group sampling is done differently as we see next.

C. Structured Populations

Real populations are highly structured. Spatial structures or social networks make some individuals interact more often than others. Recent advances in network theory [19], [20], show us that real-world populations are organized in complex interaction structures, ranging from random-like graphs to scale-free networks. Scale-free graphs [20], are of particular interest due to the fact that their degree distribution follows a power law. This means that some vertices of the graph are highly connected (often called "hubs") while the vast majority of the vertices only have a small number of connections. An example of a scale-free network is presented in figure 1A. In order to create a better model of the real world, we will also study the effects of individual learning on a PGG played on a population structured in a scale-free network.

Scale-free networks were created according to the Barabasi-Albert model of growth and preferential attachment [20]. To create a scale-free network of average degree $<\zeta>$we start

from a small number of nodes m_0, and progressively add new nodes with degree $m = <\zeta>/2 = m_0$. For there to be preferential attachment the new node connects to an existing node i with probability $p_i = \zeta_i / \sum_j \zeta_j$ where ζ_i is the degree of node i. This means that new nodes have a "preference" to attach themselves to already heavily linked nodes. By construction, the Barabasi-Albert model enforces a minimum group size of $<\zeta>/2 + 1$.

Group Sampling: Under an heterogeneous population like a scale-free network agents are no longer equally likely to play against each other. Agents now only play with their neighbors. An agent with ζ neighbors plays $\zeta + 1$ PGGs, each with a given group size. One with all neighbors and then a game for each of the neighbor's neighbors. This group sampling is represented on figure 1B. The individual fitness of the agent derives from the payoff accumulated from all games he partook in.

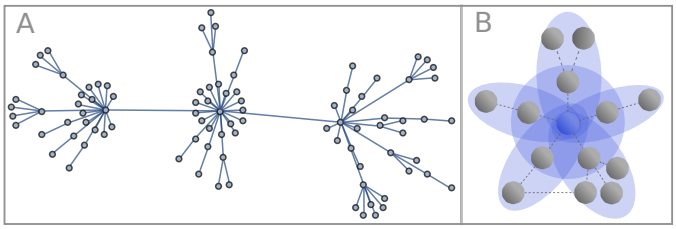


Fig. 1. (A) Example of a scale-free network. (B) Group sampling on a scale-free network: The central individual (blue) participates in 6 groups, each with its own group size. These groups are represented by the blue circles

D. Risk

Up until now, failure to reach the threshold meant the pile was not divided by the agents. With the introduction of risk ($r \in [0, 1]$) now the pile is still divided with probability $1 - r$. This means that we can consider $r = 1$ for the PGG played under traditional conditions.

E. Reinforcement Learning

Now that we know what payoffs agents get, we need to model how agents will update their strategies according to the payoffs they received. *Q-learning* [30], [32] is the technique of reinforcement learning we will mainly use. It is a so-called action value method. It consist of an update rule, an action selection rule, and an action value. The update rule determines how action values are updated based on new experience. The action selection rule determines which action to take next. The action value indicates the quality of taking one action relative to another.

Single-state Q-learning: For this algorithm all initial action values are initialized to some value $Q_0 \in \mathbb{R}$. Using small initial action values (relative to payoffs) speeds up learning in the beginning because it increases the importance of those payoffs. However, if we wish to allow for greater exploration all we have to do is use initial action values much greater than the payoffs an agent can receive. According to those action values the action selection rule will select an action to take. Actions with higher values will have more probability to be chosen than those with lower values. The action selection rule used is *softmax action selection*. We will calculate the probability $p_{s,a}$ of selecting an action a with an action value $q_{s,a}$ for current state s according to the Boltzmann distribution:

$$p_{s,a} = \frac{e^{(q_{s,a}/\tau)}}{\sum_{a'} e^{(q_{s,a'}/\tau)}} \tag{2}$$

Temperature τ controls the rate of exploration: much exploration at high temperature, little exploration at low temperature. To let the behavior stabilize, the temperature τ can be decreased over time. After taking action a and receiving payoff u the action values are updated as follows:

$$q_{s,a} \leftarrow \begin{cases} q_{s,a} + \alpha(u - q_{s,a}) & \text{If action } a \text{ was taken,} \\ q_{s,a} & \text{otherwise.} \end{cases} \tag{3}$$

where $\alpha \in [0, 1]$ is the learning rate. An higher learning rate puts more weight on more recent payoffs. If the learning rate $\alpha = 0$, nothing is ever learned; if it is $\alpha = 1$, the action value of an action simply equals the last payoff earned for that action.

F. Initial Distribution of Agents

The simulations start with an even distribution of cooperators and defectors. This can be achieved by simply creating all agents with the same action value for both strategies (cooperating and defecting). This way we would ensure all agents start with the same probability of either cooperating or defecting. However, we do not think this is a good model of the real world. Populations are not created with copies of agents that all start under the same conditions. So we chose another strategy that also ensures an even distribution of cooperators and defectors while maintaining a better degree of heterogeneity. Instead of all players choosing to cooperate with probability $p_c = 0.5$ we spread the agents' probability of cooperating across the full spectrum $p_c \in]0, 1[$. This means that some agents start with a probability of cooperating $p_c = 0.1$, others with $p_c = 0.2$, and so on. Because this is an uniform distribution the average probability of cooperating of the whole population is still $p_c = 0.5$. This was achieved by changing the initial action values of the agents.

IV. Results

A. Methods

In our model an agent is defined by its two action values (one for each behavior: cooperating or defecting). These action values represent the quality of choosing one behavior over the other. Agents play several rounds of a Public Goods Game (PGG) sequentially. In each round each agent chooses to either cooperate or defect according to its action values. Then the groups are formed and their respective payoffs calculated. In the end of the round each agent receives a certain payoff associated with its group and the behavior

chosen and updates its action values accordingly. Then, for the next round, they can choose to keep that behavior or change it. What we call a simulation is several of those rounds played sequentially. Agents keep updating their action values in every round and play the PGG for 1500 rounds per simulation. We can run several simulations (these are separate entities, not sharing any information between them) under the same conditions and then statistically analyze the results in order to have more robust data. Each point in all graphs is an average of at least 20 simulations run under the same conditions. For the scale-free networks 6 different networks were created and 100 simulations ran in each one. When the deviation between simulations was negligible it was, therefore, not shown.

First we observed the impact of changing the parameters of the Q-learning algorithm. Changing the learning rate made no difference on the amount of cooperative actions of the population after the population had stabilized. Changing the way the temperature τ of the action selection rule decreases only changes the amount of rounds needed for the agents to stabilize. After several simulations it was decided to decrease this temperature τ as such:

$$\tau(i) = min\{\tau, \tau \log(i/3)/(i/3)\} \tag{4}$$

Where $\tau(i)$ is the temperature used at the ith round and τ is a value determined by us which only needs to be high enough to make sure there is enough exploration in those first rounds. This way the agents start to stabilize at around the 600th round which gives plenty of time to explore what is the best behavior. The cost which agents pay to cooperate is fixed $c = 1$ on all subsequent graphs.

B. Influence of Group Sampling in well-mixed Populations

A set of simulations were run to see what is the effect of the stochastic processes inherent to finite populations. The simulations were run with a population of 200 agents ($Z = 200$) , group size is fixed at 5 ($N = 5$) and there is no threshold ($M = 1$). The results are shown in figure 2. Each point represents the average fraction of cooperators (after the population stabilized) for 20 simulations for a certain Fc/N. The whiskers represent the deviation of values between those 20 simulations. Each simulation was run for 1500 rounds. The fraction of cooperators was measured by taking the average of the last 500 rounds (once the population had stabilized). The red circles represent simulations where each agent participated in one and only one group. The blue squares and yellow rhombus represent simulations where each agent participated in at least 40 groups and 80 groups, respectively. For each agent on each round, the payoff received was the average of the payoffs of all the groups he participated. This way, the stochastic effects of finite populations were dampened. As we can see, those stochastic processes play a detrimental role for cooperation when $Fc/N > 1$. They also introduce more deviation between simulations. Given our goal of promoting cooperation, from now on every simulation will be run with the formation of several groups per individual. On all subsequent graphs, every agent partakes in at least 80 groups. This number was chosen because it is high enough to approach the asymptotic limit but still manageable in terms of simulation run time. Similarly to what was found analytically and with social learning [24], **on a PGG without thresholds cooperation is advantageous only when $Fc/N > 1$.**

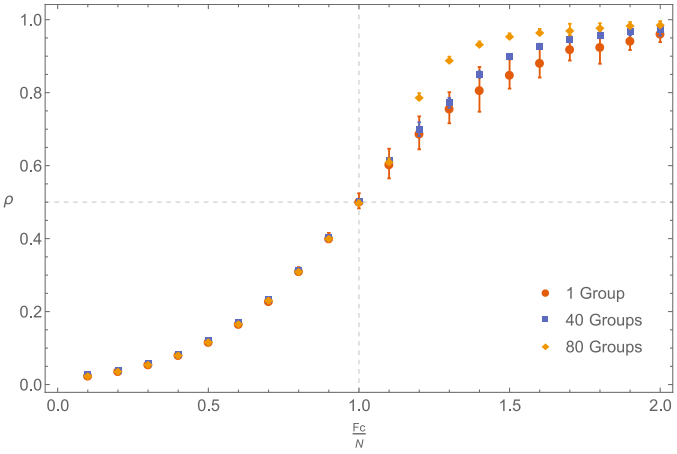


Fig. 2. Fraction of cooperative actions ρ for each value Fc/N.

C. Introduction of a Threshold

The introduction of a threshold makes the population dynamics more complex as seen in figure 3. Here the simulation parameters were similar to those of figure 2 except for the group size and the introduction of the threshold. Here group size $N = 10$. Now cooperation can be advantageous even for $Fc/N < 1$.

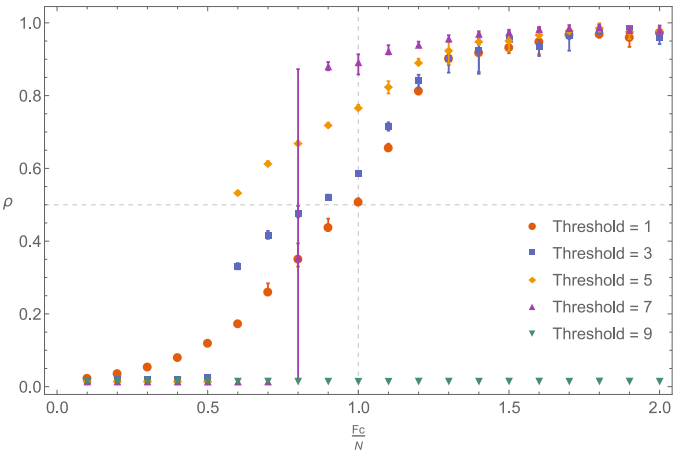


Fig. 3. Fraction of cooperative actions ρ for each value Fc/N under different Thresholds.

An interesting point appears in figure 3. For $Fc/N = 0.8$ we can see that there is an high deviation in the value of the fraction of cooperation for threshold $M = 7$. What could explain this? A look at the learning gradient might provide an answer. We use the learning gradient to understand

under which circumstances cooperating is an advantageous behavior over defecting. This is accomplished by calculating the average difference in fitness between cooperators and defectors. We multiply that difference by the fraction of cooperative actions (ρ) and defective actions ($1-\rho$) to allow an easier comparison with social learning. This way we end up with the following formula for the learning gradient:

$$L_G(\rho) = \rho(1-\rho)(f_C - f_D) \tag{5}$$

Where f_C and f_D is the average fitness (payoffs) of agents that chose to cooperate and defect in that round, respectively. This learning gradient characterizes the behavioral dynamics of the population. Whenever the gradient is positive ($L_G(\rho) > 0$) it means that a cooperative action is providing more payoff than a defective action, which means that cooperation is more likely to be reinforced by individual learning. Inversely, if the learning gradient is negative ($L_G(\rho) < 0$) it means that defection is more likely to be reinforced. This is presented in figure 4. For each curve the learning gradient is zero

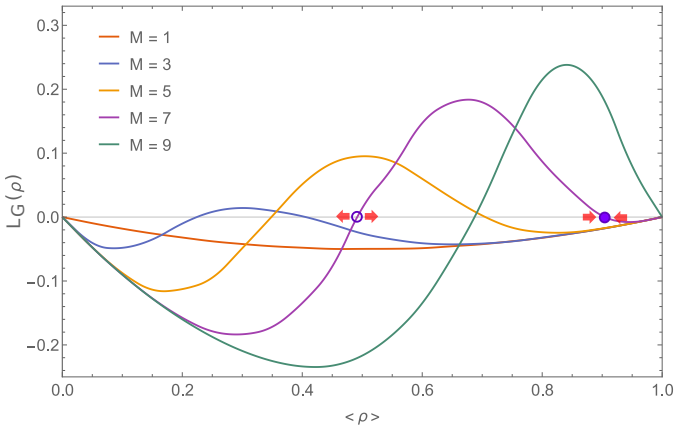


Fig. 4. Learning gradient $L_G(\rho)$ under different thresholds for $Fc/N = 0.8$. The open circle represents the unstable fixed point and the filled circle represents the stable fixed point. The arrows indicate the direction in which the amount of cooperative actions will tend to move for threshold $M = 7$.

$L_G(\rho) = 0$ in two points. The one to the left is an unstable fixed point and the one to the right is a stable fixed point. On all simulations, initially, the population is evenly divided between agents that choose to defect or cooperate. Figure 4 shows that for threshold $M = 7$ and a fraction of cooperative actions $\rho = 0.5$ the learning gradient is 0. This is the unstable fixed point. For this reason, depending on whether the random nature of the game favors cooperators or defectors, the population can either enter a state of full defection or a state where the fraction of cooperators is around $\rho = 0.85$. The latter point is the stable fixed point. For $\rho = 0.85$ and $M = 7$ any deviation in the composition of the population produced by the stochastic processes is negated after some rounds and the population remains with the same fraction of cooperators. The unstable fixed point explains the high deviation present in figure 3 for point $Fc/N = 0.8$ and $M = 7$. The nature of these fixed points leads us to the following conclusion: **as long as there are enough cooperative actions in the population to surpass the unstable fixed point of the threshold, its existence is advantageous for cooperation.** This results are, once again, similar to what was found under social learning [24].

D. Group Sizes

So far the group size was always much smaller than the population size ($N \leq 10$ and $Z \geq 200$). What happens if we increase the group size to levels close to those of the population size? As with social learning [12], [24], once the group size spans nearly the entire population we can observe the "spite" effect [25], which is detrimental for cooperation. This is shown in figure 5. The simulations were run with a population of 500 agents ($Z = 500$). The enhancement factor was Fc/N = 0.8. **Cooperation is maximized when groups are small.** This is valid both when the threshold was constant or increased linearly with the group size (however with the latter, to a lesser extent).

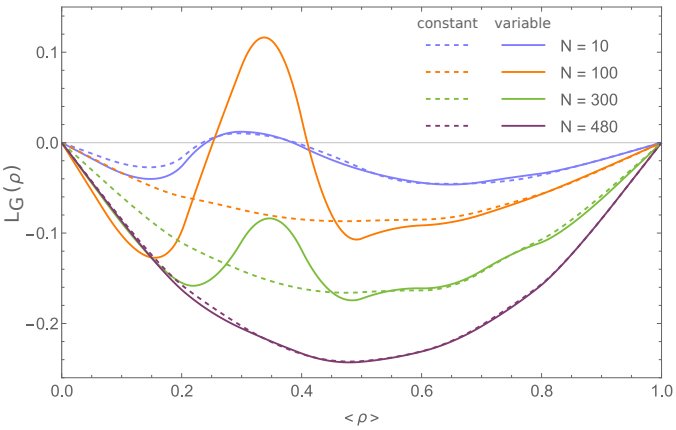


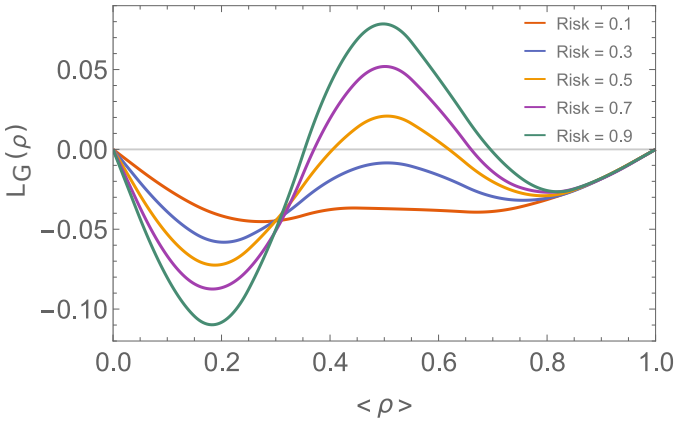
Fig. 5. Learning gradient $L_G(\rho)$ for several group sizes (N). On all dashed curves the threshold is constant $M = 3$. Solid curves represent simulations where the threshold increases linearly with group size $M = 0.3N$.

E. Risk

Following what was said about risk (section III-D) figure 6 shows the results of introducing risk in the PGG. Group size $N = 10$ and threshold $M = 5$. Once again, by analyzing the stable fixed point in figure 6B (the rightmost point where $L_G(\rho) = 0$) we can see that as risk gets higher so does the fraction of cooperators present in the population. The results present in figure 5 and 6 further reinstate the findings of Santos and Pacheco [12]: *"When applied to the problem of climate control, the present results suggest that decentralized agreements between smaller groups (small N), possibly focused on region-specific issues where risk is high and goal achievement involves tough requirements (large relative M), may be preferable to world summits, because they effectively raise the probability of reaching an overall cooperative state."*

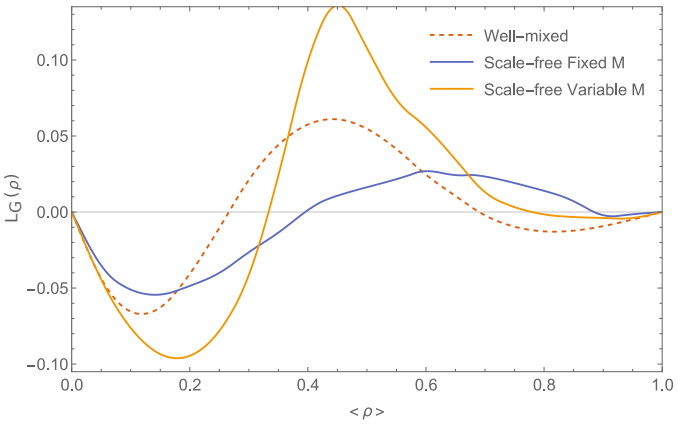
F. Structured Populations

With all the previous results, we can have a good understanding of the behavioral dynamics of the PGG played by an unstructured well-mixed population. But how will the network



structure affect the behavioral dynamics of the population? Figure 7 shows the changes on the learning gradient with structured populations. The dashed line represents the well-mixed population (as we were using previously) and acts as a baseline. The blue and yellow curves represent the structured population (scale-free). The conditions of the simulations under the well-mixed and structured populations were equivalent:

- Population size $Z = 1000$ agents
- Group size $N = 7$ for well-mixed population. Each agent participates in 7 groups per round instead of 80 as was being used before. The scale-free networks were created with an average degree $\zeta = 6$ which means the average group size is $<N>= 7 = <\zeta>+1$. It also means that, on average, each agent will participate in 7 groups per round.
- The threshold is constant $M = 3$ for the well-mixed population and for the blue curve. For the yellow curve the threshold increases linearly with the group size $M = 3N/7$. The choice of$M = 3N/7$ ensures that the average value of$M$ is the same for both curves.
- Enhancement factor is $Fc/N = 0.9$.



Looking at figure 7, we can observe that **cooperators take advantage of such an heterogeneous network**. The stable fixed point (the rightmost point where $L_G(\rho) = 0$) is higher in both structured populations as compared with the well-mixed one. This is especially true for the structured population where the threshold is fixed. However, we can also observe that even though the stable fixed point is higher for the blue curve, its height on the y-axis is much smaller. This means that this type of structure with a fixed threshold is only advantageous under stricter circumstances than its variable threshold counterpart. This is observable, for example, if the enhancement factor is $Fc/N = 0.8$. In that case, the scale-free network with variable threshold is still advantageous over the well-mixed network. However, the learning gradient for the structured network with a fixed threshold is always negative, meaning that cooperators are worse off than with the well-mixed structure.

V. Conclusion

In this work we presented individual learning as another tool to study the emergence of cooperation. We mainly focused on computational simulations of agents using reinforcement learning playing a Public Goods Game (PGG). These simulations gave us a good insight of how selfish, rational agents behave when presented with the choice of incurring a cost to themselves to give a benefit to another agent. **As conflicting as these ideas may seem, it was shown that selfish agents can still choose to cooperate, under certain circumstances.**

A. Summary of Contributions

Proposal of new model: There has been a lot of research effort trying to explain the emergence of cooperation using social learning. In this work we tackled the problem in a different fashion: what would happen if agents were actually trying to learn by themselves, instead of imitating each other's behavior? In this thesis we propose a model where agents play a PGG while using reinforcement learning in an attempt to maximize their payoffs.

The influence of group sampling: The agents form groups from the population to play the PGG. The more groups an agent partakes in, the better for cooperation. Participating in more groups dampened the negative stochastic effects of finite populations. It also improved the final payoffs of the agents that chose to cooperate as compared to those who did not.

Introduction of a Threshold: By introducing a threshold in the PGG this game no longer has a linear return. Now if a group fails to meet that threshold, all agents of that group receive no payoff. This provides a further incentive to cooperate. The higher the threshold the more cooperators stand to gain, as long as there are enough cooperators in the population to surpass the unstable fixed point.

Group sizes: Cooperation is maximized when the groups playing the PGG are much smaller than the whole population.

Introduction of Risk: If we introduce risk in the PGG (on failure to reach the threshold, agents still receive the payoff with probability $1 - r$). Using our model it was shown that the higher the risk the more agents are willing to cooperate.

Structured Populations: Real populations are highly structured. Instead of agents interacting equally likely to each other, a structured population network was introduced. It was shown that cooperators take advantage of the heterogeneous nature of the scale-free network to a great extent.

All results obtained with individual learning further reinstate the findings made with social learning. It seems that these two mechanisms are alike, although they work at different levels of organization.

B. Future Work

It is hard to say that a work is complete. There are several possible paths to enhance the present solution, and some are presented next.

More learning algorithms: We only used *Q-learning* in our model. For the results present here to be more robust it would be important to have other learning algorithms produce similar results. One could use, for example, *Roth-Erev learning* [29] and see how those agents fared compared to those using *Q-learning*.

Experimental Validation: It would be interesting to try to mimic the experimental results of Milinski et al. [13] with reinforcement learning. If the experimental results using real people were the same as with agents using reinforcement learning we could make the case that maybe humans behave like selfish, rational agents. This could provide some insight into our own human nature.

Impact of Networks: We studied briefly the impact of scale-free networks on the behavioral dynamics. We saw that hubs tend to mainly want to cooperate while the other nodes fall to either side of the spectrum. But a more thorough analysis can follow. What made hubs behave so differently than the other nodes?

Individual learning and Social Learning: Even though individual learning produced similar results to those of social learning, it is not clear what would be the results of both learning processes employed simultaneously. This approach could provide a more accurate model of human behavior.

REFERENCES

[1] M. A. Nowak, "Five rules for the evolution of cooperation," *science*, vol. 314, no. 5805, pp. 1560–1563, 2006.

[2] P. E. Stander, "Cooperative hunting in lions: the role of the individual," *Behavioral Ecology and Sociobiology*, vol. 29, no. 6, pp. 445–454, 1992.

[3] C. Boesch, "Cooperative hunting roles among tai chimpanzees," *Human Nature*, vol. 13, no. 1, pp. 27–46, 2002.

[4] S. Creel and N. M. Creel, "Communal hunting and pack size in african wild dogs, lycaon pictus," *Animal Behaviour*, vol. 50, no. 5, pp. 1325–1339, 1995.

[5] R. L. Trivers, "The evolution of reciprocal altruism," *Quarterly review of biology*, pp. 35–57, 1971.

[6] B. Beding, "The stone-age whale hunters who kill with their bare hands." daily mail. see http://www.dailymail.co.uk/news/article-465987/the-stone-age-whale-hunters-kill-bare-hands.html 4 jul. 2007." Web, 2010, accessed: 2015-12-10.

[7] R. Jervis, "Cooperation under the security dilemma," *World politics*, vol. 30, no. 02, pp. 167–214, 1978.

[8] J. Bryant, *Coordination theory, the stag hunt and macroeconomics*. Springer, 1994.

[9] R. E. Michod, "Mediation during the origin of multicellularity," *Genetic and cultural evolution of cooperation*, p. 291, 2003.

[10] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior*. Princeton university press, 2007.

[11] R. M. Dawes, "Social dilemmas," *Annual review of psychology*, vol. 31, no. 1, pp. 169–193, 1980.

[12] F. C. Santos and J. M. Pacheco, "Risk of collective failure provides an escape from the tragedy of the commons," *Proceedings of the National Academy of Sciences*, vol. 108, no. 26, pp. 10 421–10 425, 2011.

[13] M. Milinski, R. D. Sommerfeld, H.-J. Krambeck, F. A. Reed, and J. Marotzke, "The collective-risk social dilemma and the prevention of simulated dangerous climate change," *Proceedings of the National Academy of Sciences*, vol. 105, no. 7, pp. 2291–2294, 2008.

[14] G. Hardin, "The tragedy of the commons," *science*, vol. 162, no. 3859, pp. 1243–1248, 1968.

[15] W. Hamilton, "The genetical evolution of social behaviour. i," 1964.

[16] M. A. Nowak and K. Sigmund, "Evolution of indirect reciprocity by image scoring," *Nature*, vol. 393, no. 6685, pp. 573–577, 1998.

[17] E. Lieberman, C. Hauert, and M. A. Nowak, "Evolutionary dynamics on graphs," *Nature*, vol. 433, no. 7023, pp. 312–316, 2005.

[18] M. A. Nowak and R. M. May, "Evolutionary games and spatial chaos," *Nature*, vol. 359, no. 6398, pp. 826–829, 1992.

[19] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[20] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[21] F. Santos, J. Rodrigues, and J. Pacheco, "Graph topology plays a determinant role in the evolution of cooperation," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 273, no. 1582, pp. 51–55, 2006.

[22] A. Traulsen and M. A. Nowak, "Evolution of cooperation by multilevel selection," *Proceedings of the National Academy of Sciences*, vol. 103, no. 29, pp. 10 952–10 955, 2006.

[23] A. Traulsen, M. A. Nowak, and J. M. Pacheco, "Stochastic dynamics of invasion and fixation," *Physical Review E*, vol. 74, no. 1, p. 011909, 2006.

[24] J. M. Pacheco, F. C. Santos, M. O. Souza, and B. Skyrms, "Evolutionary dynamics of collective action in n-person stag hunt dilemmas," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 276, no. 1655, pp. 315–321, 2009.

[25] W. D. Hamilton, "Selfish and spiteful behaviour in an evolutionary model," 1970.

[26] M. O. Souza, J. M. Pacheco, and F. C. Santos, "Evolution of cooperation under n-person snowdrift games," *Journal of Theoretical Biology*, vol. 260, no. 4, pp. 581–588, 2009.

[27] F. C. Santos, M. D. Santos, and J. M. Pacheco, "Social diversity promotes the emergence of cooperation in public goods games," *Nature*, vol. 454, no. 7201, pp. 213–216, 2008.

[28] L. R. Izquierdo, S. S. Izquierdo, and F. Vega-Redondo, "Learning and evolutionary game theory," in *Encyclopedia of the Sciences of Learning*. Springer, 2012, pp. 1782–1788.

[29] A. E. Roth and I. Erev, "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games and economic behavior*, vol. 8, no. 1, pp. 164–212, 1995.

[30] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, University of Cambridge England, 1989.

[31] E. Kalai and E. Lehrer, "Rational learning leads to nash equilibrium," *Econometrica: Journal of the Econometric Society*, pp. 1019–1045, 1993.

[32] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," *AAAI/IAAI*, no. s 746, p. 752, 1998.